23: ANOVA in R

Stat 120 | Fall 2025

Prof Amanda Luby

0.1 ANOVA in R

When we are performing an ANOVA test for a difference in means, we can use the `aov()` function in R. Here is a template:

```
aov(y ~ x, data = dataset)
```

where

- `y` is the column name for the response variable
- `x` is the column name for the categorical predictor variable
- `dataset` is the name of the data set

1 Wordsum

The code below reads in a dataset from the General Social Survey that includes scores on a 10-question vocabulary test called "wordsum" and a self-reported class (one of "lower", "working", "middle" or "upper").

The vocabulary test works as follows: respondents are given a list of 10 words, and are asked to choose a word from the list that comes closest to the meaning of the first word provided in the capital letters.

- SPACE (school, noon, captain, room, board, don't know)
- BROADEN (efface, make level, elapse, embroider, widen, don't know)
- EMANATE (populate, free, prominent, rival, come, don't know)
- EDIBLE (auspicious, eligible, fit to eat, sagacious, able to speak, don't know)
- ANIMOSITY (hatred, animation, disobedience, diversity, friendship, don't know)
- PACT (puissance, remonstrance, agreement, skillet, pressure, don't know)
- CLOISTERED (miniature, bunched, arched, malady, secluded, don't know)
- CAPRICE (value, a star, grimace, whim, inducement, don't know)

- ACCUSTOM (disappoint, customary, encounter, get used to, business, don't know)
- ALLUSION (reference, dream, eulogy, illusion, aria, don't know)

```
gss = read_csv("https://raw.githubusercontent.com/OpenIntroStat/ims-tutorials/master/05-infer/08
gss
```

```
# A tibble: 795 x 2
   wordsum class
     <dbl> <chr>
 1       6 MIDDLE
 2       9 WORKING
 3       6 WORKING
 4       5 WORKING
 5       6 WORKING
 6       6 WORKING
 7       8 MIDDLE
 8      10 WORKING
 9       8 WORKING
10       9 UPPER
# i 785 more rows
```

We're interested in whether scores on the vocabulary test vary based on self-reported `class` (middle, working, lower, or upper).

1. What is the *response variable* and what is the *predictor variable*
2. Make an appropriate EDA to start to answer this question
3. Write out an appropriate hypothesis test
4. Calculate the test statistic using the `aov()` function. Some starter code is below (you will need to remove the line that says `eval = FALSE`)

```
wordsum_aov = aov(<response> ~ <predictor>, data = gss)
summary(_______)
```

4. Check assumptions using the code below. Remove the line that says `eval = FALSE`

```
p1 = ggplot(wordsum_aov, aes(x = .fitted, y = .resid))+
  geom_point()
p2 = ggplot(wordsum_aov, aes(x = .resid)) +
  geom_histogram(bins = 15, col = "white")
p1+p2
```

5. Report your p-value and interpret results